Shubham Agarwal

+91 9083271307 skejriwal44@gmail.com Github google-scholar LinkedIn webpage

Research Interest

My research interest lies at the intersection of **ML and Systems**. I focus on co-designing algorithms and systems for large-scale training and deployment of generative models. I have experience in building systems at Adobe and have published works leveraging techniques such as approximate computing, constraint-based scheduling, and runtime adaptation in distributed environments.

Education

Bachelor of Engineering – Computer Science and Engineering GPA: 9.95/10.00
Birla Institute of Technology and Science (BITS) Pilani, Pilani Campus, India 2018 – 2022
Scored the highest cumulative GPA among 1035 students over four years
Undergraduate Thesis Project: Multi-Modal Deep Learning for Automated Document Processing Workflows

Research Experience

Research Associate – Adobe Research, India Jul 2022 – present
Systems and Insights Group. Collaborators: Dr. Subrata Mitra, Dr. Shiv K Saini

- Designed a high throughput inference serving system for text-to-image models using a novel caching technique.
- Developed an outage forecasting and root cause diagnosis pipeline for microservice-based architectures on the cloud.
- Published **6 papers**, filed **3 patents,** and successfully integrated research technologies into **2 products**.

Research Intern – American Express, India Jan 2022 – Jul 2022
Document Analytics & Intelligence Lab. Mentor: Dr. Himanshu S Bhatt

- Designed a multi-modal deep learning pipeline for processing visually rich documents using YOLO and LayoutLM.
- Developed a scalable framework for swift adaptation to new document types through automated few-shot learning.
- Implemented cloud-based inferencing using PyTorch models, and on-device inference using lightweight TFLite models.

Research Intern – Adobe Research, India May 2021 – Aug 2021
Big Data Intelligence Lab. Mentor: Dr. Shiv K Saini

- Used cloud observability data to predict outages, considering the rarity of these events in real-world data.
- Developed a framework to model outages as extreme events by forecasting the distribution of QoS metrics and using a tail-risk regularizer for precise modeling of the distribution tails.
- Resulted in a full paper **publication** at FSE 2023 and a **patent** filed with the USPTO.

Publications

[1] **S Agarwal**, S Mitra, S Chakraborty, S Karanam, K Mukherjee, and S Saini. "Approximate Caching for Efficiently Serving Text-to-Image Diffusion Models". In *The 21st USENIX Symposium on Networked Systems Design and Implementation*. NSDI 2024. (*acceptance rate 18%*)

[2] G S Ahmad, **S Agarwal**, S Mitra, R A Rossi, M Doshi, and S Paila. "ScaleViz: Scaling Visualization Recommendation Models on Large Data". In *The 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. PAKDD 2024.

[3] **S Agarwal**, S Chakraborty, S Garg, S Bisht, C Jain, A Gonuguntla, and S Saini. "Outage-Watch: Early Prediction of Outages using Extreme Event Regularizer". In *The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. FSE 2023. (*acceptance rate 25.6%*)

[4] S Chakraborty, **S Agarwal**, S Garg, A Sethia, U Pandey, V Aggarwal, and S Saini. "ESRO: Experience Assisted Service Reliability against Outages". In *The 38th IEEE/ACM International Conference on Automated Software Engineering*. ASE 2023. (*acceptance rate 21.3%*)

[5] S Chakraborty, S Garg*, **S Agarwal***, A Chauhan, and S Saini. "CausIL: Causal Graph for Instance Level Microservice Data". In *Proceedings of The Web Conference*. WWW 2023. (*acceptance rate 19.2%*)

[6] **S Agarwal**, G Chan, S Garg, T Yu, and S Mitra. "Fast Natural Language Based Data Exploration with Samples". In *Companion of the 2023 International Conference on Management of Data*. SIGMOD 2023 (*Demo*).

Pre-prints: * *equal contribution*

[1] **S Agarwal**, S Iqbal, and S Mitra. Micro-managing Prompts for High-Throughput Text-to-Image Inference Serving using Approximation. (*under review at ATC 2024*)

[2] C Lu*, **S Agarwal***, M Tanjim, K Mahadik, A Rao, S Mitra, S Saini, S Bagchi, S Chaterji. RECON: Accelerating Text-to-Image Diffusion with Retrieval of Concept Prompt Trajectories. (*under review at ECCV 2024*)

[3] A Ikram, K Lee, **S Agarwal**, S Mitra, S Saini, S Bagchi, and M Kocaoglu. FRC-EG: Algorithm for Finding Root Cause of Failures in Microservices Efficiently via Causal Structure. (*under review*)

Patents:

[1] **S Agarwal**, S Mitra, S Chakraborty, S Karanam, K Mukherjee, S Saini. Intelligent Use of Caching and Retrieval of Intermediate Noise for Resource Efficient Diffusion Models. [Being filed with USPTO]

[2] **S Agarwal**, S Mitra, R A Rossi, G S Ahmad, M Doshi, and S Paila. Reinforcement Learning Based Framework for Scaling Visualization Recommendation Models on Large Data. [Being filed with USPTO]

[3] **S Agarwal**, G Chan, S Garg, T Yu, S Mitra. Interactive Sequential Data Exploration using NLP with Sampling based Approximations. [Being filed with USPTO]

[4] S Garg, **S Agarwal**, S Bisht, N Sheoran, C Jain, A Gonuguntla, S Saini. A System and Method for Outage Forecasting. (US Patent App. 17/656,263)

Selected Projects

System for ML *Papers published in* [NSDI '24, PAKDD '24]

- **Approximate Caching for Diffusion Models (In Product)**: Developed a caching-based text-to-image pipeline to reduce inference latency by up to 2×. Deployed the system on a GPU cluster that uses a Vector Database server for cache indexing and EFS for cache storage. Used a CPU instance to pipeline the database and storage calls alongside GPU execution.
- **High-Throughput Text-to-Image Inference Serving**: Co-designed the resource allocator and query scheduler of an inference serving system for diffusion model on a fixed size multi-node GPU cluster. Reduces latency SLO violations by 10× and offers 10% higher average quality of generation and 40% higher throughput by employing *accuracy-scaling*.
- **Scaling VizRec Models on Large Data**: Designed a plug-in framework for Visualization Recommender systems that achieves up to 10× speedup in latency. Implemented scalable Deep Q-Learning to optimize input statistics selection.

ML for System Reliability *Papers published in* [FSE '24, ASE '24]

- **Outage Prediction in Cloud**: Designed an outage forecasting model using distribution learning to reduce the mean time to detect outages by up to 88% for a large-scale microservice-based deployment on the cloud. Developed a method to retrieve real-time monitoring metrics from Prometheus for inference, while also supporting re-training of the model.
- **Root Cause and Remediation Consolidation System**: Developed a diagnostic service for cloud failures that merges both alert data and incident reports and achieves 27% improvement in root cause recommendations.
- **Runtime Prediction of Incoming Jobs (In Product)**: Built a pipeline to train and predict Spark job latency within a multi-tenant data processing platform. Ingested job metadata from Kafka and used Azure Functions to scale the deployment.

Research Collaborations

- **On-Device Speculative Decoding and Routing for LLMs** Mar 2024 – Present
  with Prof. Hui Guan, University of Massachusetts, Amherst
  - Exploring *accuracy-scaling* to optimize the on-device inference efficiency of LLMs.
- **Training-free acceleration for text-to-image synthesis** Oct 2023 – Present
  with Prof. Somali Chaterji, Purdue University
  - Designed a novel concept retrieval-based method for accelerating text-to-image models.*(paper under review)*
- **Causality-based Root Cause Diagnosis for microservice architectures** Jul 2023 – Present
  with Prof. Saurabh Bagchi, CRISP Lab, Purdue University
  - Developed an essential graph-based causal inference algorithm for RCA in complex systems.*(paper under review)*

Technical Skills

- Programming languages: Python, C, C++, Java, SQL, Verilog, MIPS
- Packages & Frameworks: PyTorch, Keras, TensorFlow, scikit-learn, Git, Docker, Kubernetes, Kafka

Undergraduate Courses:

Operating Systems, Database Systems, Computer Networks, Compilers, Computer Architecture, Artificial Intelligence, Information Retrieval, Data Structures and Algorithms, Probability and Statistics, Linear Algebra

Course Projects

- [GitHub] An enhanced collaborative filtering recommender system with fine-tuned item weights and similarity scores.
- [GitHub] An AI tutoring system for teaching algebra, generating questions based on a dynamic reward function.
- [GitHub] A custom grammar parser and compiler in C for language design and type expression computation.
- [GitHub] A REACT app for seamless online education, integrating content sharing and video calling features.
- [GitHub] A lightweight progressive web app client using JavaScript to work with the metastudio.org server.

Activities & Achievements

- Industry mentor for UMass Master's students' group, guiding research on On-device model security.
- Mentored 10 undergraduate interns and collaborated with 2 PhD interns during summer internships at Adobe.
- Contributed to multiple paper reviews for technical conferences including ATC, AISTATS, AAAI, EACL, etc.
- Awarded 100% merit scholarship for academic excellence across all semesters; ranking top among all students.
- Served as Student Senate Representative for over 1000 undergraduate students in the Student Faculty Council.
- Successfully organized BITS Pilani's Annual Technical Fest, boosting external participation and outreach by 200%.